MATH. SCAND. 6 (1958), 175–180

TWO REMARKS ON THE BASIC THEOREMS OF INFORMATION THEORY

LENNART CARLESON

1. In a recent paper by A. J. Khinchin [1], Shannon's basic theorems on information theory have been given their first thorough mathematical treatment. In his paper, Khinchin critizises the definitions of the concept of channel capacity given earlier and shows that the converses of Shannon's encoding theorems are not obvious—in Khinchin's paper no such converses are given. The purpose of this note is to show that, although the critizism is justified, the definition given by Khinchin is actually equivalent to the standard definition; this result will then immediately give the desired converses[1]. First, however, we shall in section 2 outline the basic concepts and show that the Shannon-McMillan theorem holds for any denumerable probability field with finite entropy.

2. Let $p_1, p_2, \ldots, p_n, \ldots$ be the probabilities belonging to a given denumerable probability field F. The entropy of F is defined by (logarithms are taken to the base 2)

$$H(F) = -\sum_1^\infty p_\nu \log p_\nu$$

and is assumed to be finite. Let $x = \{x_n\}_{-\infty}^{\infty}$ be a stationary stochastic process, where x_n assumes e.g. only the values $1, 2, \ldots$, with probabilities $p_1, p_2, \ldots$. Let $\mu(x)$ be the probability distribution on the product space Ω corresponding to this process and denote by C_n the cylindric subset of Ω determined by a given set of values of $x_1, \ldots, x_n$. The entropy of the multivariate distribution of the C_n's is given by

$$H_n = -\sum_{C_n} \mu(C_n) \log \mu(C_n) ,$$

and it is easy to see that $\lim_{n\to\infty}(H_n/n) = H$ exists. For our purposes it is important that even

Received September 1, 1958.

[1] After this paper was submitted, the author became aware that this result was proved (1958) in I. P. Zaregradski, *Eine Bemerkung über die Durchlasskapazität eines stationären Kanals mit endlichem Gedächtnis*, Arbeiten zur Informationstheorie II, Berlin 1958. The present proof, based on the relation (2.1) below, is however considerably simpler.

$$\lim_{n\to\infty}(H_{n+1}-H_n) = H \tag{2.1}$$

exists (cf. Shannon [2, Theorem 5]).

To prove this (and more), we write

$$H_{n+1}-H_n = \sum_{C_{n+1}} \log \frac{\mu(C_n)}{\mu(C_{n+1})}\cdot\mu(C_{n+1}) = \int_\Omega g_n(x)\,d\mu(x)\,,$$

where

$$g_n(x) = \log\frac{\mu(C_n)}{\mu(C_{n+1})}, \qquad \mu(C_0) = 1, \quad n = 0, 1, \ldots .$$

The functions $g_n(x)$ are non-negative and it was shown in [1] that $\lim_{n\to\infty} g_n(T^{-n}x) = g(x)$ exists almost everywhere, T being the shift transformation. With obvious modifications, this proof applies also in our denumerable case. If we can show that $g(x)$ is summable and

$$\lim_{n\to\infty}\int_\Omega g_n(x)\,d\mu(x) = \int_\Omega g(x)\,d\mu(x)\,, \tag{2.2}$$

the proof of (2.1) is complete, since

$$H = \lim_{n\to\infty}\frac{H_n}{n} = \lim_{n\to\infty}\frac{1}{n}\sum_{\nu=0}^{n-1}\int_\Omega g_\nu(x)\,d\mu(x)\,.$$

This now follows from the following argument (Hilfssatz 7.3 in [1]).

Let $E_{n,k}$ denote the subsets of Ω where

$$\begin{aligned} E_{n,k}&: \quad 2^k \leqq g_n(x) < 2^{k+1}, \quad k = 1, 2, \ldots;\\ E_{n,0}&: \quad 0 \leqq g_n(x) < 2, \quad k = 0\,. \end{aligned}$$

Let Z_ν denote the set with $x_{n+1}=\nu$; we shall assume, as we may, that $\mu(Z_\nu)=p_\nu$ is a non-increasing sequence. For $x\in Z_\nu$, $C_{n+1}=C_n\cap Z_\nu$ and if x also belongs to $E_{n,k}$ we have

$$\mu(C_n\cap Z_\nu) \leqq 2^{-2^k}\mu(C_n)\,. \tag{2.3}$$

Hence

$$\begin{aligned}\sum_{k=K}^{\infty}\int_{E_{n,k}} g_n(x)\,d\mu(x) &= \sum_{\substack{k\geqq K\\ \nu\geqq 1}}\int_{E_{n,k}\cap Z_\nu} g_n(x)\,d\mu(x)\\ &\leqq \sum_{\substack{k\geqq K\\ \nu\geqq 1\\ C_n}} 2^{k+1}\,\mu(C_n\cap E_{n,k}\cap Z_\nu)\\ &= \sum_{\log\nu<2^{k-1}} + \sum_{\log\nu\geqq 2^{k-1}} = \Sigma_1+\Sigma_2\,.\end{aligned}$$

In the first sum we use the inequality (2.3) and find

$$\Sigma_1 \leqq \sum_{k \geqq K} 2^{-2^{k-1}} \cdot 2^{k+1} \sum_{C_n} \mu(C_n) = \sum_{k \geqq K} 2^{-2^{k-1}} 2^{k+1} .$$

For Σ_2 we have the estimate

$$\begin{aligned} \Sigma_2 &\leqq \sum_{k \geqq K} 2^{k+1} \sum_{\log \nu \geqq 2^{k-1}} \mu(Z_\nu) \\ &= \sum_{\log \nu \geqq 2^{K-1}} p_\nu \sum_{k \leqq 1+\log\log \nu} 2^{k+1} \\ &\leqq 8 \sum_{\log \nu \geqq 2^{K-1}} p_\nu \log \nu . \end{aligned}$$

Since $\sum_1^\infty p_\nu = 1$ and p_ν decreases it follows that $\nu p_\nu \leqq 1$ whence

$$\sum_1^\infty p_\nu \log \nu \leqq \sum_1^\infty p_\nu \log \frac{1}{p_\nu} < \infty .$$

The estimates of Σ_1 and Σ_2 depend only on K and we conclude that

$$\int_{|g_n| > \lambda} g_n(x)\, d\mu(x) < \varepsilon(\lambda) ,$$

where $\varepsilon(\lambda) \to 0$, $\lambda \to \infty$. By standard theorems on Lebesgue integrals, the statement (2.2) follows.—We finally remark that the McMillan theorem for our case now follows as in [1].

3. We shall now consider transmission of information in that we introduce a second stationary process $y = \{y_n\}$, called the output. We suppose that x_n as well as y_n can only assume A resp. B different values. We furthermore assume that y depends on x in the following way. Let Z be an arbitrary cylindric set in the y-space, corresponding to given values of certain y_i, $k \leqq i \leqq l$. The conditional probability $\nu_x(Z)$ is then assumed to depend only on x_j, $k-m \leqq j \leqq l$, where m is a fixed integer (the "memory" of the "channel" defined by ν_x)[1]. The joint process (x, y) is also stationary. We denote the corresponding entropies by $H(x)$, $H(y)$ and $H(x, y)$, respectively. The fundamental concept in the theory is the capacity C of the channel,

$$C = \sup_\mu h(\mu, \nu_x) , \tag{3.1}$$

where

$$h(\mu, \nu_x) = H(x) + H(y) - H(x, y) . \tag{3.2}$$

[1] In [1], this is assumed only for $Z = \{y \mid y_i = b\}$. However, in the proof, e.g. on p. 64, the above more general assumption is used.

In Khinchin's paper it is pointed out that in taking the upper bound we must restrict ourselves to ergodic μ's, that is, to sources for which every translation invariant set E of x's, $E=TE$, has probability 0 or 1. We shall here prove that this restriction is not necessary since the upper bound C is the same whether or not we make the restriction:

THEOREM. *Let $\varepsilon>0$ be given, let ν_x be a fixed channel, and let μ be an arbitrary (not necessarily ergodic) stationary distribution on the space of x. Then there exists an ergodic μ^* such that*

$$h(\mu^*, \nu_x) \geqq h(\mu, \nu_x) - \varepsilon ,$$

where h is defined in (3.2).

PROOF. We choose N so large that

$$|H_n(z) - H_{n-1}(z) - H(z)| < \delta, \qquad n \geqq N ,$$

holds for $z=x$, y, and (x, y).

For all cylindric sets Γ_N, determined by a set of values of $x_\nu, x_{\nu+1}, \ldots, x_{\nu+N-1}$, we define

$$\mu^*(\Gamma_N) = (1-\alpha)\mu(\Gamma_N) + \alpha A^{-N}, \qquad 0 < \alpha < 1; \tag{3.3}$$

for unions E of disjoint Γ_N's, $\mu^*(E)$ is defined additively. This definition is obviously consistent, $\mu^*(E)=\mu^*(TE)$, whenever these numbers are defined, and $\mu^*(\Omega)=1$. Let Γ_{N+1} be a continuation of Γ_N by $x_{\nu+N}$ and define Γ'_N and Γ'_{N-1} by

$$\Gamma_{N+1} = \Gamma_N x_{\nu+N} = x_\nu \Gamma'_N = x_\nu \Gamma'_{N-1} x_{\nu+N} .$$

We extend the definition of μ^* by defining $(\mu^*(\Gamma'_{N-1})>0)$

$$\mu^*(\Gamma_{N+1}) = \mu^*(\Gamma_N) \cdot \frac{\mu^*(\Gamma'_N)}{\mu^*(\Gamma'_{N-1})} ;$$

for unions of disjoint Γ_{N+1}'s, μ^* is defined as above. Obviously,

$$\sum_{x_{N+\nu}=1}^{A} \mu^*(\Gamma_{N+1}) = \mu^*(\Gamma_N)$$

and

$$\sum_{x_\nu=1}^{A} \mu^*(\Gamma_{N+1}) = \mu^*(\Gamma'_{N-1}) \cdot \frac{\mu^*(\Gamma'_N)}{\mu^*(\Gamma'_{N-1})} = \mu^*(\Gamma'_N) ,$$

in agreement with our previous definition. By induction, $\mu^*(\Gamma_n)$ is defined in a consistent way for all n and is then extended to the Borel sets of Ω. The resulting process, $x^*=\{x_n^*\}$, is clearly stationary. Furthermore, if ξ_i^* denotes the vector

$$\xi_i^* = (x^*_{i(N-1)}, x^*_{i(N-1)+1}, \ldots, x^*_{i(N-1)+N-2}), \qquad i = 0, \pm 1, \ldots,$$

$\xi^* = \{\xi_i^*\}$ constitutes, by the construction, a stationary Markov process with A^{N-1} different states. Since $\alpha > 0$, all transition probabilities are positive. Hence the ξ^*-process is ergodic (cf. e.g. Doob, *Stochastic processes*, p. 460). If E_x is a translation invariant set of x^*'s, the relation $E_x = T^{N-1}E_x$ shows that the corresponding set of ξ^*'s is invariant and so $\mu^*(E_x) = 0$ or 1. The x^*-process is thus ergodic.

We shall now estimate the entropies corresponding to this input x^*. We first observe that $\{x_n^*\}$ has the entropy

$$H(x^*) = H_N(x^*) - H_{N-1}(x^*) .$$

Actually, if $n > N$, it follows from a general formula ([1, p. 31]) that

$$H_n(x^*) - H_{n-1}(x^*) = H_N(x^*) - H_{N-1}(x^*) ,$$

and the left-hand side tends to $H(x^*)$. On the other hand, $H_N(x^*)$ and $H_{N-1}(x^*)$ differ for small α arbitrarily little from $H_N(x)$ resp. $H_{N-1}(x)$, and so, by the choice of N,

$$|H(x^*) - H(x)| < 2\delta \qquad \text{for} \quad \alpha < \alpha_0 .$$

It is easy to see that the joint process (x^*, y^*), where y^* is the output corresponding to x^*, is a Markov process of order $N + m - 1$. As above, we conclude that

$$|H(x, y) - H(x^*, y^*)| < 2\delta, \qquad \alpha < \alpha_1 .$$

In order to prove a corresponding inequality for y^* we first assume $\alpha = 0$. In this case ξ^* need not be ergodic, but let it consist of the closed *ergodic* subchains $C_1, C_2, \ldots, C_k$ and let η^* be constructed as above as the vector-form (of length $N - 1 > m$, where N was defined above) of the y^*-process. We want to estimate the conditional probability $P_{\eta_1^*\eta_2^*\ldots\eta_{n-1}^*}(\eta_n^*)$ in terms of η_i^*, $n - p < i < n$. This probability is obtained by considering (1) the probability q_ν that η_{n-1}^* was obtained from a ξ_{n-1}^* in the chain C_ν, (2) the transition probabilities of C_ν and (3) the conditional probabilities $\nu_{\xi_{n-1}^*\xi_n^*}(\eta_n^*)$. Since the chains C_ν are ergodic, the probability distribution of ξ_{n-1}^* in C_ν will depend arbitrarily little on the given $\eta_1^*\eta_2^*\ldots\eta_{n-p}^*$, where p can be kept fixed as $n \to \infty$. Hence, q_ν is the only quantity in the above argument which may depend on the remote past of η^*. We have two possibilities. (a) $\eta_{n-p+1}^*, \ldots, \eta_{n-1}^*$, p fixed, $n \to \infty$, will determine a definite subchain C_ν with arbitrarily high probability, if p is large enough, in which case the desired estimate has been obtained.

(b) The output corresponding to a certain number of chains C_ν has the same distribution; in this case, the individual determination of the q_ν's corresponding to these chains is irrelevant in order to get the conditional probability of η_n^* and the estimate is obtained also in this case. We have thus proved:

$$\left| P_{\eta_1^*\eta_2^*\dots\eta_{n-1}^*}(\eta_n^*) - P_{\eta_{n-p+1}^*\dots\eta_{n-1}^*}(\eta_n^*) \right| < \delta' \tag{3.4}$$

if $p \geqq p(\delta')$, except for a set of given η_i^*'s, $1 \leqq i \leqq n-1$, of probability $< \delta'$. —The same estimate now obviously also holds as soon as $\alpha < \alpha_2$, since the transition probabilities change arbitrarily little and a transition from C_ν to C_μ, $\nu \neq \mu$, during the last p steps has arbitrarily small probability. Finally,

$$H_n(\eta^*) - H_{n-1}(\eta^*) = \sum_{\eta_1^*,\dots,\eta_{n-1}^*} P(\eta_1^*, \dots, \eta_{n-1}^*)\, H_{\eta_1^*,\dots,\eta_{n-1}^*}(\eta_n^*)\,,$$

where, in the sum, H denotes the conditional entropies of η_n^*. If we use the estimate (3.4), when it is valid, and the fact that H is uniformly bounded, we see that, as $n \to \infty$,

$$\left|\left(H_n(\eta^*) - H_{n-1}(\eta^*)\right) - \left(H_{p-1}(\eta^*) - H_{p-2}(\eta^*)\right)\right| < \tfrac{3}{2}\delta, \qquad \alpha < \alpha_3\,,$$

which at last yields

$$|H(y^*) - H(y)| < 2\delta, \qquad \alpha < \alpha_4\,.$$

If $\delta = \frac{1}{6}\varepsilon$ we find for $\alpha < \alpha_5$

$$h(\mu^*, \nu_x) \geqq h(\mu, \nu_x) - \varepsilon\,.$$

Since x^* is ergodic, this proves the theorem.

REFERENCES

1. A. J. Khinchin, *Über grundlegende Sätze der Informationstheorie*, Arbeiten zur Informationstheorie I, Berlin, 1957.
2. C. E. Shannon, *A mathematical theory of communication*, Bell System Techn. J. 27 (1948), 379–423, 623–656.

UNIVERSITY OF UPPSALA, SWEDEN